Lecture 12: Analysis of Selection Experiments

Bruce Walsh lecture notes
Synbreed course
version 4 July 2013

Variance in the Response to Selection

$R = h^2S$ is just the expected value of the response, but there is a variance about this value.

Hence, identically-selected replicate lines are Still expected to show variation is response

The major source of such variation is genetic drift

Consider the mean in generation t

$$z_t = \mu + g_t + d_t + e_t$$

μ = Mean of the original population

g_t = The mean breeding value in generation t

d_t = the effect of any major environmental trend in generation t

e_t = error in estimating the environmental-corrected mean breeding value from the mean phenotype of a sample

Under this model, the mean of a replicate series of lines is

$$E(z_t) = \mu + E(g_t) + d_t$$

$R = h^2S$

Consider the mean in generation t

$$E(z_t) = \mu + E(g_t) + d_t$$

The variance is given by

$$\sigma_e^2(t) = \frac{\sigma_z^2}{M_t}$$

$$\sigma_z^2(t) = \sigma_g^2(t) + \sigma_e^2(t) + \sigma_d^2$$

Variance in the breeding value at generation t

Variance in the environmental trend (mean is set to be zero)

In generation t, M_t individuals are measured. An upper bound on the error variance is $Var(z)/M_t$

Variance in Breeding Values

Two sources of variation

(i) Sampling variance in the founding lines

(ii) Genetic drift (inbreeding) within each line

$$\sigma_g^2(t) = \left(\frac{1}{M_0} + 2f_t\right)\sigma_A^2 = \left(\frac{1}{M_0} + 2f_t\right)h^2\sigma_z^2$$

M_0 = Size of the founding population

f_t = Inbreeding in generation t

$$2f_t = 2\left[1 - \left(1 - \frac{1}{2N_e}\right)^t\right] \simeq t/N_e \qquad \text{for } t/N_e << 1$$

The mean breeding values in different generations of the same replicate line are correlated,

$$\sigma\left(g_t, g_{t'}\right) = \sigma\left(z_t, z_{t'}\right) = \left(\frac{1}{M_0} + 2f_t\right)h^2\sigma_z^2 \quad \text{for } t < t'$$

Variance-covariance structure within a line

Assume the initial sample is sufficiently large so that we can ignore $1/M_0$

Variance: $\sigma^2(g_t) = (t/N_e)h^2\sigma^2_z$

Covariance: $\sigma^2(g_t, g_x) = (x/N_e)h^2\sigma^2_z$ for $x < t$

These expressions (which are often called the pure-drift approximations) will prove useful in the statistical analysis of selection response

The Realized Heritability

Since R = h^2 S, this suggests h^2 = R/S, so that the ratio of the observed response over the observed differential provides an estimate of the heritability, the realized heritability

Obvious definition for a single generation of response. What about for multiple generations of response?

Cumulative selection response = sum of all responses

$$R_C(t) = \sum_{i=1}^{t} R(i)$$

Cumulative selection differential = sum of the S's

$$S_C(t) = \sum_{i=1}^{t} S(i)$$

(1) The Ratio Estimator for realized heritability

= total response/total differential,

$$\widehat{h}_r^2 = \frac{R_C(T)}{S_C(T)}$$

(2) The Regression Estimator --- the slope of the regression of cumulative response on cumulative differential

$$R_C(t) = h_r^2 S_C(t) + e_t$$

Regression passes through the origin (R = 0 when S = 0). Slope =

$$\widehat{b} = \frac{\sum_t S_C(t) R_C(t)}{\sum_t S_C(t)^2}$$

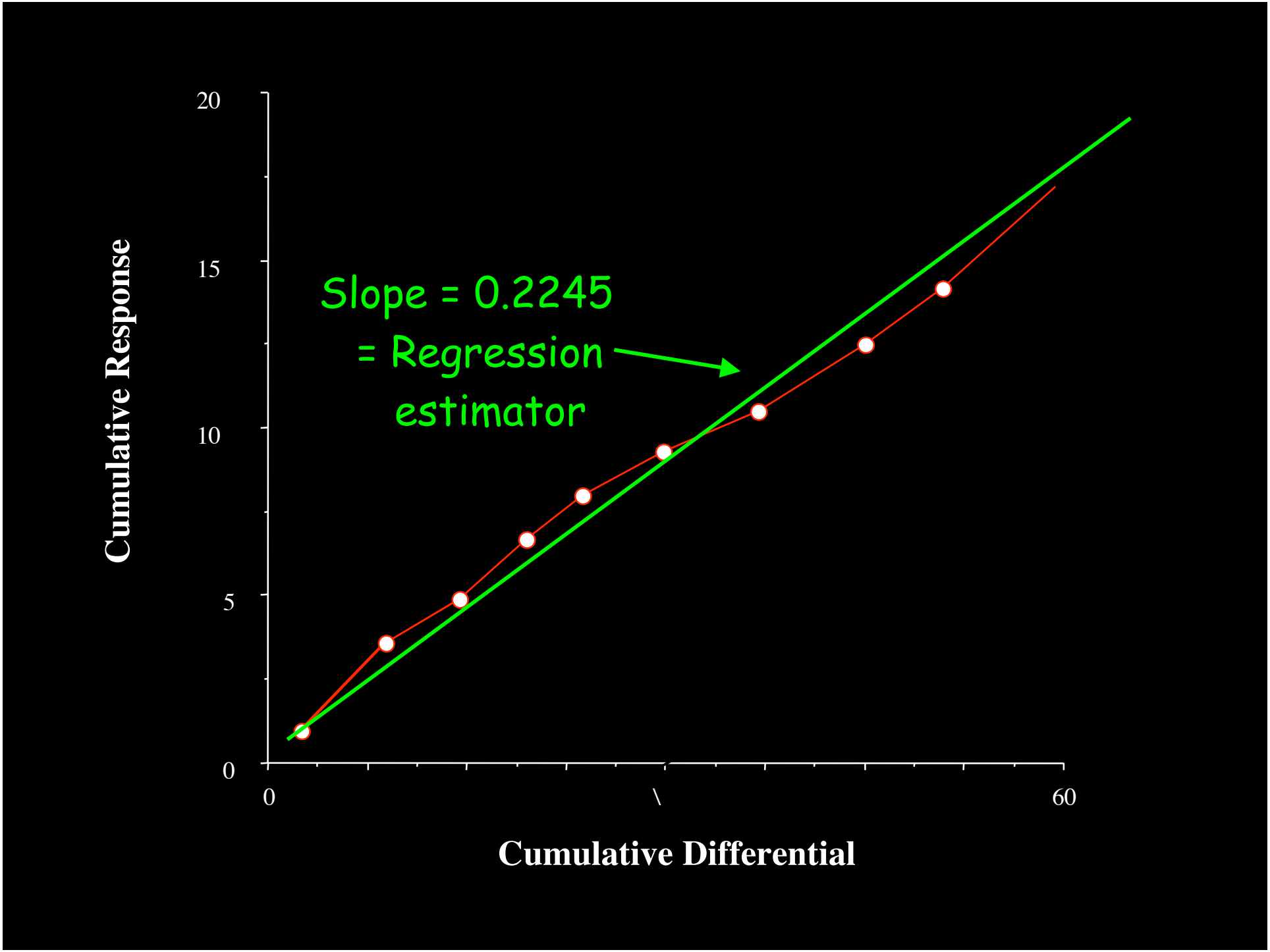
Example: Divergent selection on abdominal bristle number in Drosophila (data from T. MacKay)

t	$\overline{z}$	$\overline{z}^*$	$S(t)$	$R(t)$	$S_C(t)$	$R_C(t)$
1	18.02	20.10	$20.10-18.02 = 2.08$	$18.34-18.02 = 0.32$	.08	0.32
2	18.34	21.00	$21.00-18.34 = 2.66$	$19.05-18.34 = 0.71$	.74	1.03
3	19.05	21.75	$21.75-19.05 = 2.70$	$20.07-19.05 = 1.02$	.44	2.05
4	20.07	22.55	$22.55-20.07 = 2.48$	$20.36-20.07 = 0.29$	.92	2.34
5	20.36	22.95	$22.95-20.36 = 2.59$	$20.65-20.36 = 0.29$	12.51	2.63
6	20.65					

Ratio estimate: $h^2 = 2.63/12.51 = 0.2102$

Regression (OLS) estimator

$$\widehat{h}_r^2 = \widehat{b}_C(OLS) = \frac{\sum_{i=1}^5 S_C(i) \cdot R_C(i)}{\sum_{i=1}^5 S_C^2(i)} = \frac{78.96}{350.45} = 0.2245$$


Cumulative Response
20
15
10
5
0
Slope = 0.2245
= Regression
estimator
0
\
60
Cumulative Differential

Standard error of the Ratio Estimator

Ratio Estimator, $h^2_r = R^T/S^T$

Recall that the variance for the mean in generation t is $\sigma^2(g_t) + \sigma^2(e) + \sigma^2(d)$

Assume $M_0 >> 1$ and that we can ignore the environmental trend variance, then

$$\sigma^2(R^T) = \sigma^2(g_t) + \sigma^2(e) = (T/N_e)h^2\sigma^2_z + \sigma^2_z/M_T$$

$$\sigma^2(R^T/S^T) = \sigma^2(R^T)/(S^T)^2$$

This follows since $\sigma^2(ax) = a^2\sigma^2(x)$

Hence, $\sigma^2(h^2_r) = [\ (T/N_e)h^2\sigma^2_z + \sigma^2_z/M_T\ ]/(S^T)^2$

SE for (OLS) Regression Estimator

The basic linear model is

$$R_C(t) = h_r^2 \, S_C(t) + e_t$$

X = S_c

β = b_c

y = R

$$\widehat{b}_C(\mathrm{OLS}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\,\mathbf{y} = \left(\mathbf{S}^T\mathbf{S}\right)^{-1}\mathbf{S}^T\,\mathbf{R} \quad = \frac{\sum_{i=1}^{T} S_C(i)\cdot R_C(i)}{\sum_{i=1}^{T} S_C^2(i)}$$

Under the OLS framework (residuals homoscedastic and uncorrelated), the linear model has the design matrix X just the vector S_c of cumulative differential and y = R

SE for (OLS) Regression Estimator

$$\mathrm{Var}\left[\widehat{b}_C(\mathrm{OLS})\right] = \sigma_e^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \sigma_e^2 \left(\mathbf{S}^T\mathbf{S}\right)^{-1}$$

$$= \sigma_e^2 \Big/ \sum_{i=1}^{T} S_C^2(i)$$

$$\widehat{\sigma}_e^2 = \frac{1}{T-1}\sum_{i=1}^{T} \widehat{e}_i^2 = \frac{1}{T-1}\sum_{i=1}^{T} \Big(R_C(i) - \widehat{h}_r^2\, S_C(i) \Big)^2$$

Problems with OLS regression approach

Although the OLS regression estimator for realized heritability is very widely used, it has fatal problems

OLS assumes the residuals are homoscedastic and uncorrelated. In reality, the covariance structure is

$$\sigma^2(e_i) = (i/N_e)h^2\sigma^2_z + \sigma^2_z/M_i$$

$$\sigma^2(e_k, e_i) = (i/N_e)h^2\sigma^2_z \text{ for } i < k$$

Hence, the GLS regression is more appropriate

The OLS gives unbiased estimates of the realized heritability, but it seriously underestimates its SE

GLS regression Estimate

$$R_C(t) = h_r^2 S_C(t) + e_t$$

$X = S_c$

$\beta = b_c$

$y = R$

The variance-covariance matrix V has elements

$V_{ii} = (i/N_e)\, h^2\, \sigma^2_z + \sigma^2_z / M_i$

$V_{ji} = V_{ij} = (i/N_e)\, h^2\, \sigma^2_z$ for $i < j$

$$\widehat{b}_C(\mathrm{GLS}) = (\mathbf{S}^T \mathbf{V}^{-1} \mathbf{S})^{-1} \mathbf{S}^T \mathbf{V}^{-1} \mathbf{R}$$

We can directly estimate the phenotypic variance from the data

h^2 is what we are trying to estimate. Use an iterative approach. Try some initial value, use GLS to update this value, use the new value for next round of updating. Continue until values stabilize.

Example: GLS estimator for MacKay's data

Here M = 100, N = 20, while Var(z) = 3.293 Building up the covariance matrix V

$$\sigma^2\,[\,R_C(i)\,] = \left(\frac{i}{N}\right) h^2\sigma_z^2 + \frac{\sigma_z^2}{M} = i{*}h^2{*}\,0.1647 + 0.03292$$

$$\sigma\,[\,R_C(i), R_C(j)\,] = \left(\frac{i}{N}\right) h^2\sigma_z^2 = i{*}h^2{*}\,0.1647 \qquad \text{for } i < j$$

$$\mathbf{V} = 0.1647{\cdot}\begin{pmatrix} h^2+0.2 & h^2 & h^2 & h^2 & h^2 \\ h^2 & 2h^2+0.2 & 2h^2 & 2h^2 & 2h^2 \\ h^2 & 2h^2 & 3h^2+0.2 & 3h^2 & 3h^2 \\ h^2 & 2h^2 & 3h^2 & 4h^2+0.2 & 4h^2 \\ h^2 & 2h^2 & 3h^2 & 4h^2 & 5h^2+0.2 \end{pmatrix}$$

Start with some initial value for h^2, update V, obtain new estimate. Repeat until convergence. Start with $h^2 = 0.21$

$$\widehat{h}_r^2 = \widehat{b}_C(GLS)^{(1)} = (\mathbf{S}^T\mathbf{V}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{V}^{-1} = 0.222197$$

Using this update, next estimate is 0.222135, which remains unchanged

Standard errors of the three estimates

For the OLS regression,

$$\mathrm{Var}\left[\widehat{b}_C(\mathrm{OLS})\right] = \sigma_e^2 \Big/ \sum_{i=1}^{T} S_C^2(i)$$ = 0.0228/350.45 = 0.0000649

$$\widehat{\sigma}_e^2 = \frac{1}{T-1}\sum_{i=1}^{T}\widehat{e}_i^2 = \frac{1}{T-1}\sum_{i=1}^{T}\left(R_C(i) - \widehat{h}_r^2\, S_C(i)\right)^2$$ = 0.091/4 = 0.0228

Hence, SE for OLS estimator is 0.0081

For the ratio estimator,

$\sigma^2(h^2_r) = [\ (T/N_e)h^2\sigma^2_z + \sigma^2_z/M_T\ ]/(S^T)^2$

= [(5/20)*0.21*3.292 + 0.03292]/12.51² = 0.00132

Giving a SE of 0.0363

Standard errors of the three estimates

Finally, for the GLS regression,

$(S^T V^{-1} S)^{-1} = 1/790.4 = 0.001265$

Hence, SE for GLS estimator is 0.0356

Summarizing:

OLS: $h^2 = 0.2245 \pm 0.0081$ Much to small!

Ratio: $h^2 = 0.2102 \pm 0.0363$

GLS: $h^2 = 0.2221 \pm 0.0356$

Just how well does the breeder's equation work?

Sheridan (1988) compared realized heritability estimates with estimates of heritability obtained from resemblances between relatives in the base populations

Punch-line: Good, but not great, fit in many settings

Problems with a wider meta-analysis is that standard errors are often not presented nor is the data presented in a form that allows their calculation.

Comparison of realized and (relative-based) Heritability estimates

Species	Significant Differences	NS difference	Total
Drosophila	14 (23%)	47 (77%)	61
Tribolium	7 (27%)	19 (73%)	26
Mice/Rats	6 (18%)	28 (82%)	34
Poultry/Quail	5 (45%)	6 (55%)	11
Swine/Sheep	8 (53%)	7 (47%)	15

Lots of ways for Breeder's equation to fail

- One: Inheritance model is wrong, and infinitesimal model (large number of loci, each of small effect) is incorrect
  - However, even with just a single major gene, BE is pretty good
- Two: Additional important features not accounted by BE

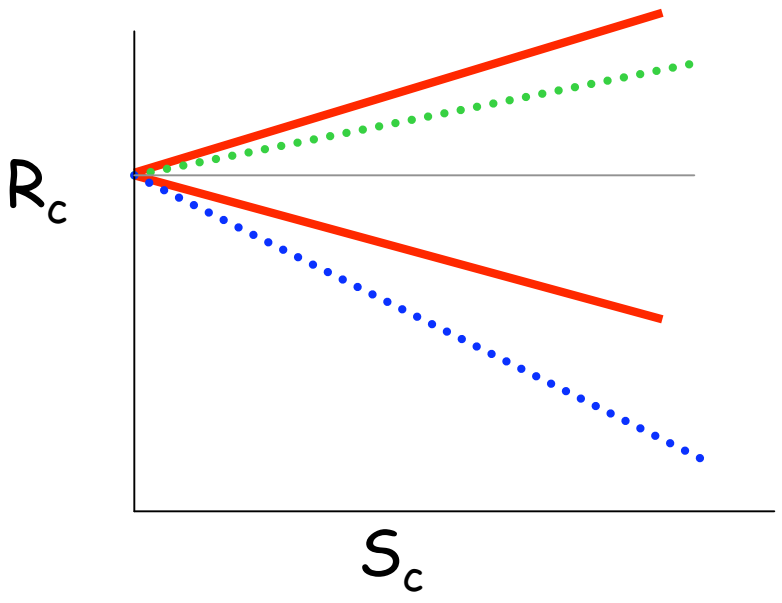
Table 13.2. Summary of various factors complicating the prediction of short-term selection response in the phenotypic mean, even assuming all regressions are linear and we are considering a single generation of selection from an unselected base population. Short-term response specifically refers to conditions where the effects of any allele frequency change on the additive variance are negligible. Models of long-term response (Chapters 26 – 28) relax this restriction.

Major gene with dominance (LW Chapter 17)	Can generate a nonlinear parent-offspring regression.
Epistasis (Chapter 15)	Component of response due to epistasis is transient. Parent-offspring covariance overestimates permanent response.
Correlated environmental effects (Chapter 15)	Contribution from parent-offspring correlation decays away after selection relaxed.
Maternal effects (Chapter 15)	Potential for complicated lags in response — mean changes unpredictably after selection is relaxed. Possibility of reversed response.
Gametic-phase disequilibrium (Chapter 16)	Changes the additive genetic variance. Directional selection generates negative gametic-phase disequilibrium, reducing h^2 and slowing response.
Assortative Mating (Chapter 16)	Generates gametic-phase disequilibrium which either enhances (positive correlation between mates) or retards (negative correlation between mates) response.

Environmental Change (Chapters 18 - 20)	A significant change in the environment can obscure the true amount of genetic change.
Drift (Chapters 18, 19)	Generates variance in the short-term response.
Environmental Correlations (Chapter 20)	Enviromental factors can influence both the trait and fitness, confounding both the nature of selection and the true amount of genetic change.
Associative effects (Chapter 22)	Trait influenced by both direct and social components from group members. A decline in the mean social value can swamp an increase in mean direct value. Possibility of reversed response.
Inbreeding (Chapter 23)	Response depends on additional variance components that are difficult to estimate (σ^2_{DI}, σ_{ADI}, etc). Response has permanent and transient components.
Selection on Correlated Characters (Volume 3)	Response completely unpredictable unless selection on correlated characters accounted for. Possibility of reversed response.
G $\times$ E Interactions (Volume 3)	Possibility of nonlinear parent-offspring regressions. Correlated characters problem, with traits measured in different environments treated as correlated traits. Possibility of reversed response.
Age-structure (Volume 3)	Several generations are required to propagate genetic change uniformly through the population.

Asymmetric Selection Response

Divergent Selection Experiment: Select some replicate lines for increased trait value, others for decreased value



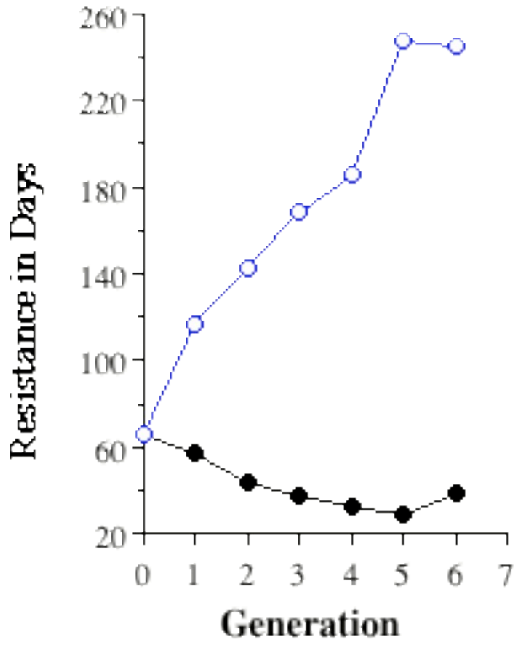
Expectation: roughly equal response in up and down directions, $R = h^2S$

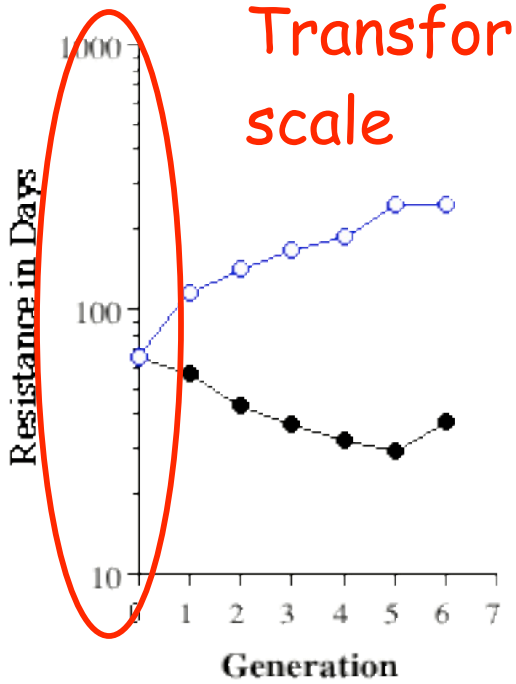
Often an asymmetric response is observed, with a significant difference in the slope of up vs. down-selection lines

Potential Causes: I. Design Defects

- Different selection differentials (Plot is R_c vs. t, not the correct plot of R_c vs. S_c)
- Drift (sample size not sufficiently large)
- Scale effects
- Undetected environmental trends
- Transient effects from previous selection
  - Decay of epistatic response
- Undetected selection on correlated traits

Scale effects

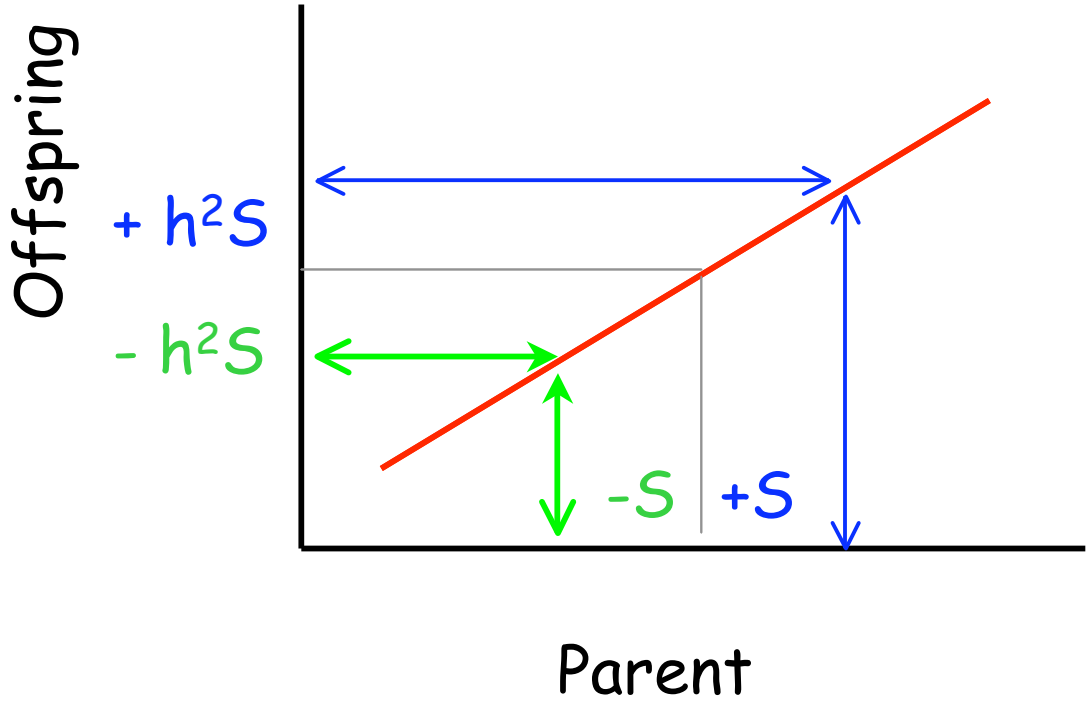




Transform to a log scale

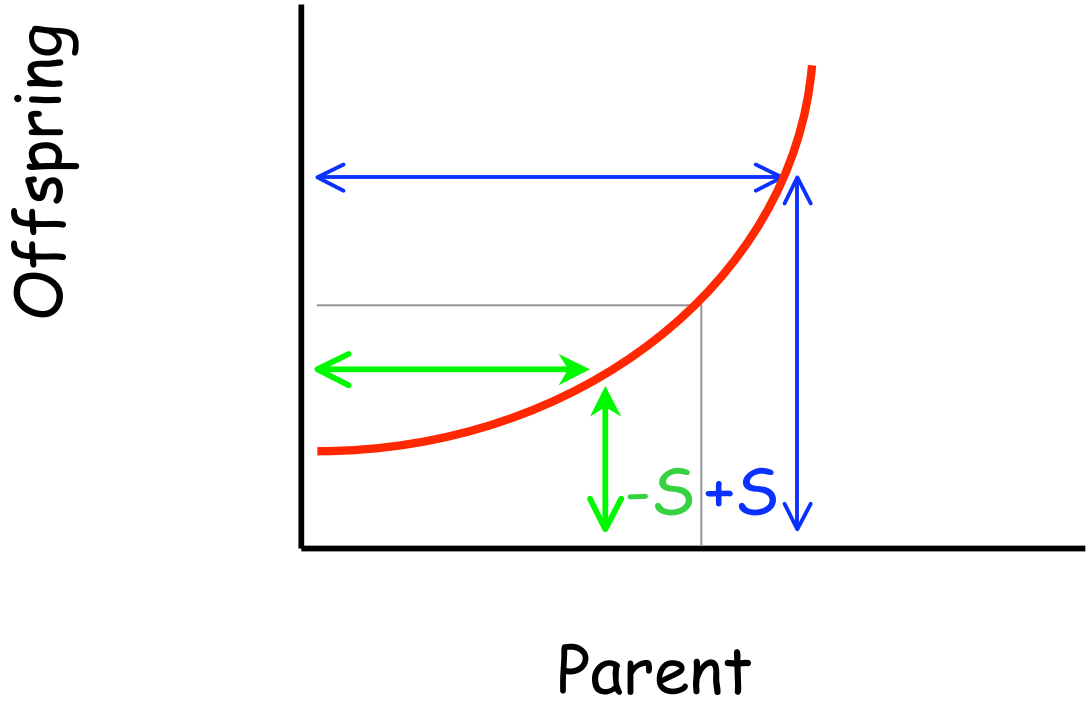
When the trait biologically cannot go below a specific value (i.e., 0), as we down-select towards zero, expect less response.

Potential Causes: II. Nonlinear Parent-Offspring regression



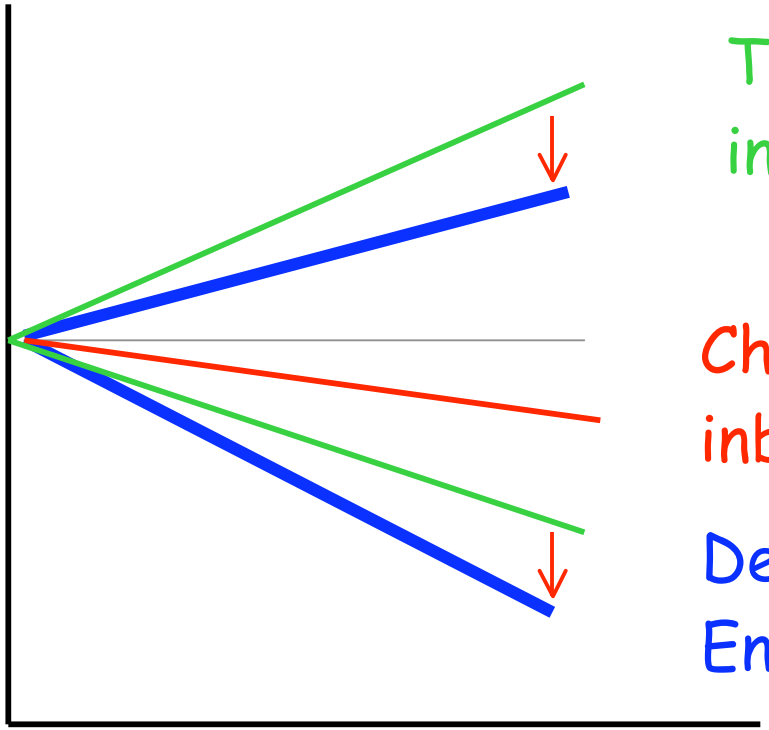
Linearity gives a symmetric response with +S, -S

However, if PO regression is nonlinear



With this nonlinear regression, larger absolute response for +S than for -S

Potential Causes: III. Inbreeding depression



True genetic response
in the absence of inbreeding

Change in mean due to
inbreeding depression.

Depresses upward response,
Enhances downward response

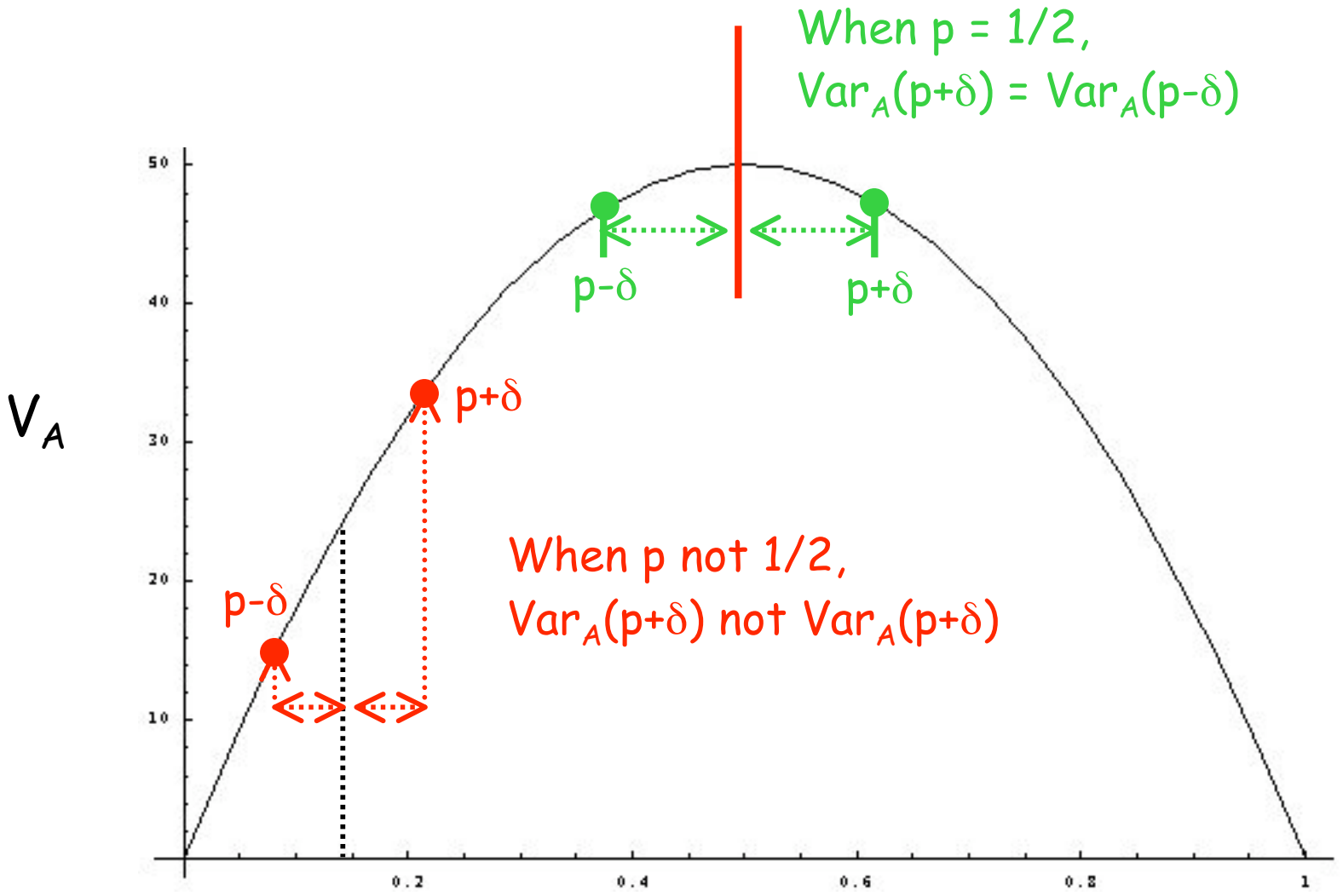
Potential Causes: IV. Genetic Asymmetry

- Requires changes in allele frequencies.

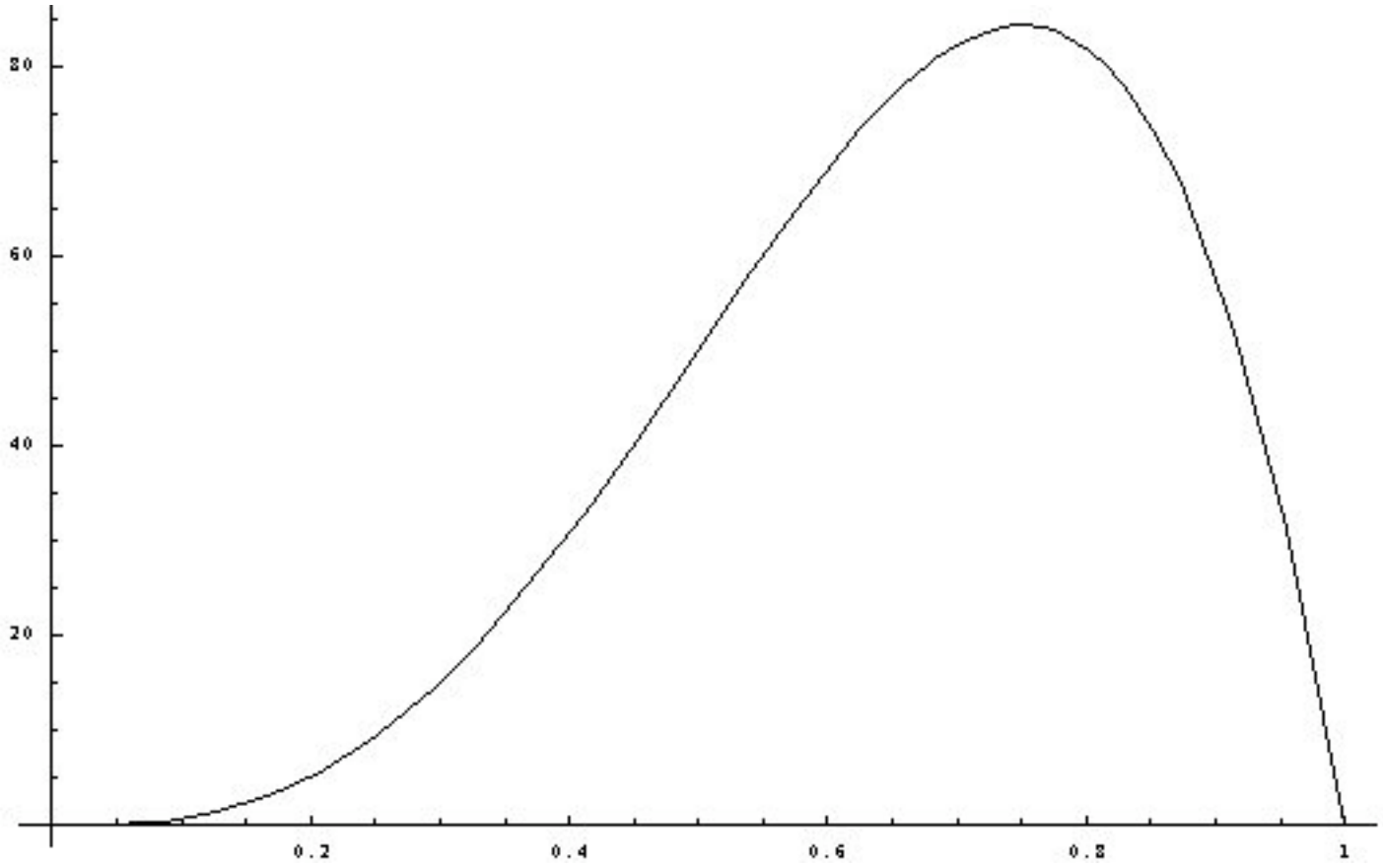
- The same absolute change in an allele frequency can result in rather different changes in the variance in the + vs. - change direction.

- This results in departures in the additive genetic variance in up vs. down-selected lines, and hence changes in h^2 and response.

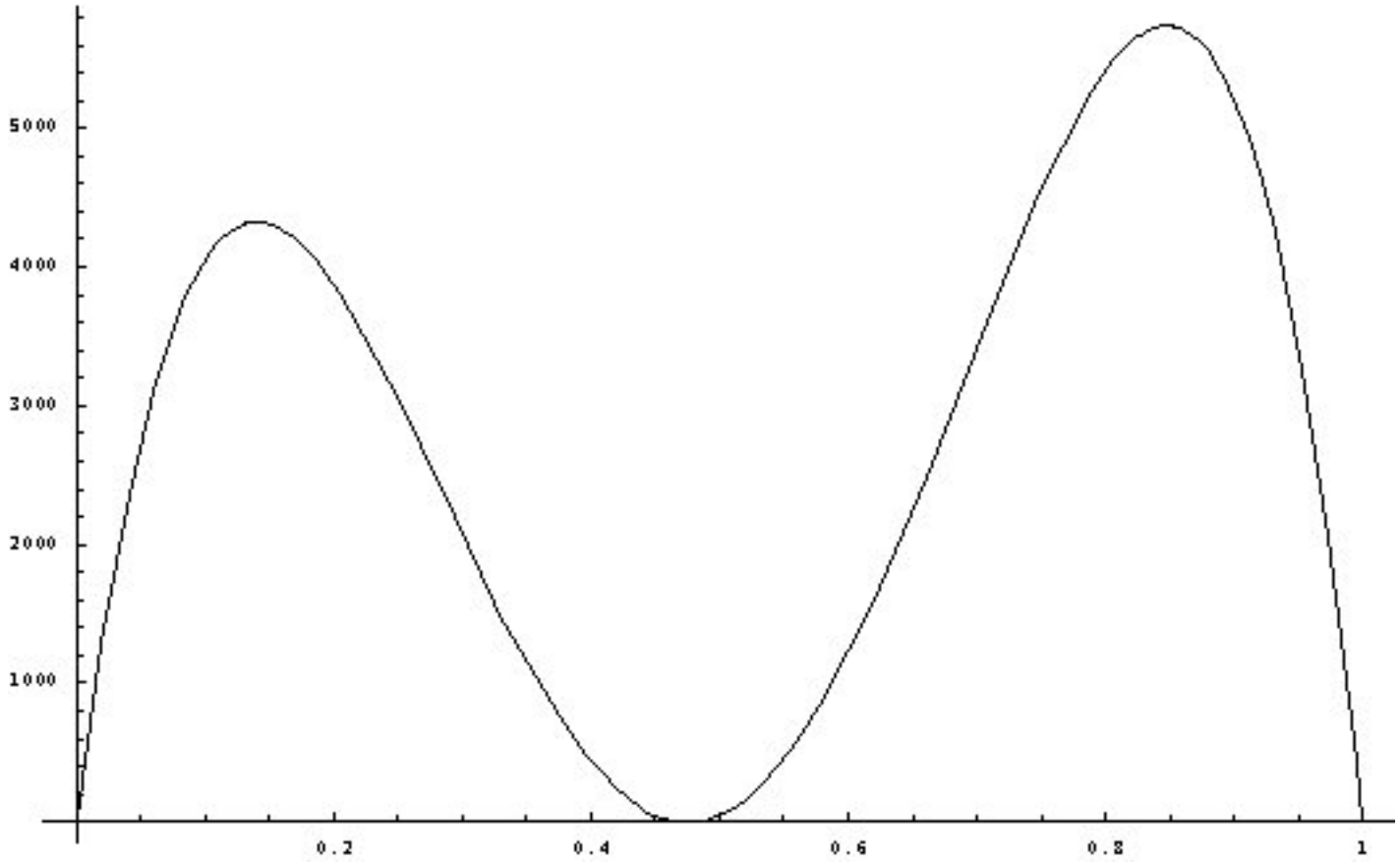
Additive variance, V_A, with no dominance (k = 0)



Additive variance, V_A, with complete dominance (k = 1)



Additive variance, V_A, with overdominance (k = 10)



Control Populations

Until now, we have been ignoring the bias caused by not accounting for any environmental trend.

One way to deal with this is to include an unselected control population in the design

$$z_{s,t} = \mu + g_{s,t} + d_t + e_{s,t}$$

$$z_{c,t} = \mu + g_{c,t} + d_t + e_{c,t}$$

Hence,

$$E(\, z_{s,t} - z_{c,t}\,) = E(\, g_{s,t}\,) - E(\, g_{c,0}\,) = h^2 S_C(t)$$

Estimating trends with a control population

$$R_t = (\, z_{s,t} - z_{c,t}) - (\, z_{s,t-1} - \; z_{c,t-1})$$

$$R_C(t) = z_{s,t} - z_{c,t}$$

$$S_t = z'_{s,t-1} - z_{s,t-1}$$

The use of a control also accounts for inbreeding depression

Complication 1: If G x E is present, then

$$E(\, z_{s,t} - z_{c,t}\,) = h^2 S_C(t) + (d_{s,t} - d_{c,t}\,)$$

Complication 2: Selection results in faster inbreeding, so control must to comparatively inbred to fully account for inbreeding depression

Divergent Selection Designs

An alternative experimental design to remove a common environmental trend is the divergent selection design

$$z_{u,t} = \mu + g_{u,t} + d_t + e_{u,t}$$

$$z_{d,t} = \mu + g_{d,t} + d_t + e_{d,t}$$

Response estimated by

$$R_t = (\, z_{u,t} - z_{u,t-1}) - (\, z_{d,t} - z_{d,t-1})$$

$$R_C(t) = z_{u,t} - z_{d,t}$$

$$S_t = (\, z'_{u,t-1} - z_{u,t-1}) - (\, z'_{d,t-1} - z_{d,t-1})$$

Note that this design also accounts for inbreeding depression (assuming up/down lines equally inbred)

Variance in Response

We have been assuming that we can ignore σ^2_d.

With a control line and/or divergent selection, don't have to worry about this.

Control:

$$R_C(t) = z_{s,t} - z_{c,t} = (\mu + g_{s,t} + d_t + e_{s,t}) - (\mu + g_{c,t} + d_t + e_{c,t})$$
$$= g_{s,t} - g_{c,t} + e_{s,t} - e_{c,t}$$

The common d_t term cancels

Divergent design

$$R_C(t) = z_{su,t} - z_{sd,t} = g_{su,t} - g_{sd,t} + e_{su,t} - e_{sd,t}$$

Again, common d_t term cancels

The resulting variance and covariances in response become

$$\sigma^2 [\, R_C(t) \,] = (2 f_t + B_0)\, 2 f_t\, h^2 \sigma_z^2 + B_t\, \sigma_z^2$$
$$\simeq (t\, A + B_0)\, h^2 \sigma_z^2 + B_t\, \sigma_z^2$$

$$\sigma [\, R_C(t), R_C(t') \,] = (2 f_t + B_0)\; h^2 \sigma_z^2$$
$$\simeq (t\, A + B_0)\, \sigma_z^2\, h^2 \qquad \text{for } t < t'$$

Design	f_t	A	B (t > t')
Selection in a single direction, no control	$f_{s,t}$	$1/N_s$	$1/M_{s,t}$
Selection in a single direction, with control	$f_{s,t} + f_{c,t}$	$1/N_{s+} + 1/N_c$	$1/M_{s,t} + 1/M_{c,t}$
Divergent Selection, no control	$f_{u,t} + f_{d,t}$	$1/N_{u+} + 1/N_d$	$1/M_{u,t} + 1/M_{d,t}$

Variance with a Control

Control populations are not without a cost.

When does the use of a control population result in a reduced variance?

Variance w/ control - variance without control =

$$\sigma^2(\, R_C(t)\,) = \left(\frac{t}{N} + \frac{1}{M_0}\right) h^2\sigma_z^2 + \frac{1}{M}\,\sigma_z^2 - \sigma_d^2$$

Hence (ignoring M terms), $\dfrac{t\sigma_z^2 h^2}{N} > \sigma_d^2$

Regardless of the value of σ^2_d, if sufficient generations are used, the optimal design (in terms of giving the smallest expected variance in response) is not to use a control.

However, this approach runs the risk of an undetected directional environmental trend compromising the estimated heritability.

Optimal Experimental Design

The coefficient of variance (CV) provides one measure for comparing different designs

$$\mathrm{CV}\,[\,R_C(t)\,] = \frac{\sigma[R_C(t)\,]}{E[R_C(t)]}$$

Design	E [R(t)]	CV [R(t)]
Selection in one direction, with control	$th^2i\sigma_z$	$(2/Nt)^{1/2}/hi$
Selection in one direction, no control	$th^2i\sigma_z$	$(1/Nt)^{1/2}/hi$
Divergent Selection, no control	$2th^2i\sigma_z$	$(1/2Nt)^{1/2}/hi$

CV scales with Nt = total # over the entire experiment

Example

Suppose we plan to select the upper 5% of the population on a trait with $h^2 = 0.25$

How large must N be to give a CV of 0.01 when no control is used?

p = 0.05 --> i = 2.06. Assuming drift variance dominates σ^2_d, then

$$CV = 0.01 = (1/Nt)^{1/2}/hi = (1/Nt)^{1/2}/(0.5*2.06)$$

$$\text{or } Nt = 1/(0.01*0.5*2.06)^2 = \boxed{9426}$$

Hence, we must have at least 9,426 selected parents over the course of the experiment

Moving from LS to MM and Bayes

- Thus far, we have assumed a least-squares (LS) analysis, which only uses information from the generation means.
- When the pedigree is known, mixed-model (MM, e.g., the animal model) is much more powerful, using all of the data and easily handling unbalanced design. MM is a likelihood analysis.
- Even better, a Bayes MM analysis fully accounts for model uncertainty (e.g., using an estimated variance for BLUP).

Mixed-Model Estimation

PROVIDED that we have the full pedigree of individuals in the selection experiment, we can use mixed-model methodology (e.g., BLUP & REML)

Power: Mixed-model accounts for ALL the covariances in the sample, not just those between means in different generations, but also ALL of the covariances between related individuals.

With sufficient connectiveness (links between relatives) between generations, can estimate the genetic trend without requiring any control population

Basic model: the animal model

$$y_{ij} = \mu + a_{ij} + e_{ij} \qquad \mathbf{Var(e)} = \sigma_e^2 \mathbf{I}$$

Vectorize the data as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_t \end{pmatrix}, \qquad \text{where} \qquad \mathbf{y}_i = \begin{pmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \\ \vdots \\ \mathbf{y}_{in_i} \end{pmatrix}$$

The (simple) model becomes $\mathbf{y} = \mathbf{1}\mu + \mathbf{a} + \mathbf{e}$

$\mathbf{Var(a)} = \sigma_A^2 \mathbf{A}$ Here, $A_{ii} = (1+f_i)$, $A_{ij} = 2\Theta_{ij}$

With additional fixed effects, $\mathbf{y} = \mathbf{Xb} + \mathbf{a} + \mathbf{e}$

The estimated mean in generation k is the average of the estimated breeding values in generation k,

$$\widehat{\overline{\mathbf{a}}}_k = \frac{1}{n_k} \sum_{j=1}^{n_k} \widehat{a}_{kj}$$

Interesting complication: The BLUP estimate of a requires a prior estimate of the heritability h^2.

REML/BLUP (also called empirical BLUP). First uses the data to obtain a REML estimate of Var(A), then use this for the BLUPs

A Bayesian analysis fully accounts for the uncertainty in using an estimate for Var(A)

REML/BLUP analysis vs. LS analysis

LS analysis uses regression to estimate realized heritability.

BLUP assumes a base population heritability and then computes the genetic trend by plotting mean BV's

BLUP produces smoother estimates, as individual BVs are compared with an index based on their relatives. BV values are regressed towards the value predicted by the index, smoothing variations out

The relationship matrix A fully accounts for the effects of drift and the generation of linkage disequilibrium (assuming the infinitesimal model holds).

This occurs because even in the face of drift and selection the covariance matrix of breeding values is the product of base-population σ^2_A times the relationship matrix (under the infinitesimal model)

This independence occurs because of the nature of the breeding value regression

$$A_i = (1/2)\, A_{fi} + (1/2) A_{mi} + s_i$$

The segregation residual s (also referred to as Mendelian sampling) is the variation generated by heterozygous parents

Segregation residual, s, is the key here

Under the infinitesimal model, s_i is independent of parental breeding values, with mean zero

$$\mathrm{Var}(s_i) = (1 - \overline{f}_i)\,(\sigma^2_A / 2)$$

Where $\overline{f}_i$ is the mean inbreeding for i's parents, and σ^2_A is the base population additive variance

More generally, the vector s ~ MVN(0, $[\sigma^2_A / 2]$ F)

Where F is a diagonal matrix with *i*-th element

$$F_{ii} = (1 - \overline{f}_i) = \left(1 - \frac{f_k + f_j}{2}\right) = \left(2 - \frac{A_{kk} + A_{jj}}{2}\right)$$

Distribution of s ***unaffected by BVs of parents***, and hence selection and/or assortative mating

Going Bayesian

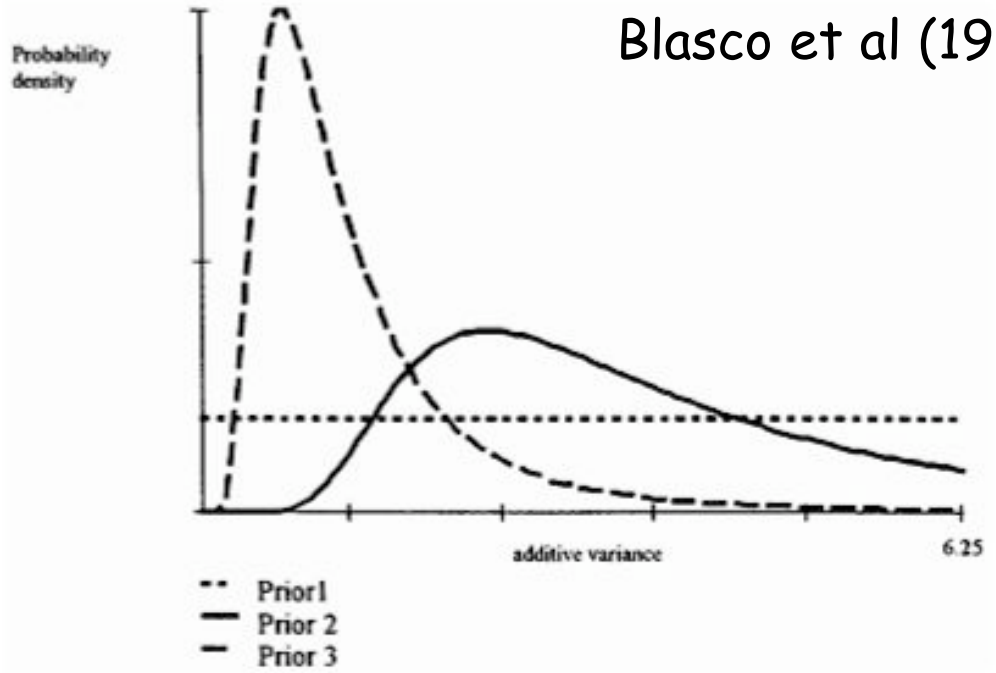
A MCMC sampler for a Bayesian mixed model is straightforward (see WL Appendix 3).

One important (but very subtle) issue is that PBVs (predicted breeding values) are used in estimate a genetic trend. The problem is that PBVs are correlated, and hence have a GLS structure. This is especially problematic with the smaller pedigrees found in wild populations. Results in mean PBV-year regressions being unbiased but highly anticonservative (p values are highly biased towards small values)

Using the distribution of slopes in an MCMC sampler avoids this problem

Example of Bayesian analysis in Large White pigs

Blasco et al (1998)



Response in Ovulation Rate at Puberty

Method	Gen 1	Gen 2	Gen 3	Gen 4
Bayesian, Prior 1	0.30 ± 0.31	0.51 ± 0.35	1.03 ± 0.39	1.58 ± 0.43
Bayesian, Prior 2	0.31 ± 0.30	0.51 ± 0.34	1.05 ± 0.38	1.55 ± 0.42
Bayesian, Prior 3	0.31 ± 0.31	0.51 ± 0.35	1.01 ± 0.35	1.53 ± 0.38
LS	-0.09	0.35	1.98	1.87
REML/BLUP	0.27	0.45	1.00	1.54

LS, MM, or Bayes?

Just what analysis should an investigator use for a selection experiment? Obviously, in the absence of any pedigree information, a least-squares analysis is the only option, although this could also be placed in a Bayesian framework. With the pedigree in hand (either observed or inferred, see Chapter 20), a mixed-model analysis is much more powerful and is strongly preferred over LS, unless there is strong evidence that model assumptions are violated. If a mixed-model approach is appropriate and chosen, should the analysis be standard or Bayesian? As mentioned, the Bayesian approach does a much better job of treating uncertainty, but this comes at a higher computational cost, especially when one does a proper analysis using several different priors to assess sensitivity. Perhaps the best advice is that offered by Blasco (2001):

> " The choice of one school or the other should be related to whether these are solutions in one school that the other does not offer, to how easily the problems are solved, and to how comfortable scientists feel with the way they convey their results. "

Blasco's last point is especially important: It is far more important for invesigators to use a method with which they are comfortable, in the sense of knowing its limitations and having some intution into the approach, than to simply use a method because it is new and trendy.

Generally speaking, simpler methods (such as OLS) tend to be more robust to model fragility than more complex approaches (e.g., mixed-models). While the latter can be considerably more powerful *when* model assumptions hold, they can also be significantly more biased when they fail. Best practice is s to use several different approaches in the analysis of any dataset. If the results are consistent, one had additional confidence the model assumptions may be holding. If they yield rather different results, this is critical for the investigator to know, suggesting a much more careful examination of model assumptions may be in order.